基于Hadoop的微博舆情监控系统模型研究

杨爱东　刘东苏
(西安电子科技大学经济与管理学院　西安　710126)

摘要:【目的】针对当前的大数据环境，提出基于 Hadoop 的微博舆情监控系统模型，实现对海量微博信息的采集、挖掘、监控分析。【方法】分析舆情监控技术，构建舆情监控系统模型，改进相关算法，利用 Hadoop 搭建大数据平台，进行仿真实验，验证模型可用性。【结果】实验结果表明，模型能够很好地对海量微博数据进行监控分析，达到舆情监控的目的。【局限】Hadoop 集群规模较小；没有对比多种聚类算法，未得到改进算法与其他算法的优劣。【结论】该模型可以对海量微博数据进行舆情监控分析，为决策者应对舆情危机提供科学化的信息支持。

关键词: 舆情监控　Hadoop　微博　大数据
分类号: G350

1　引　言

随着互联网的快速发展，Internet 已经成为当今时代信息传播的主要渠道，也是舆情[1]传播的重要途径。中国互联网络信息中心(CNNIC)第 36 次《中国互联网络发展状况统计报告》显示，截至 2015 年 6 月，中国网民规模 6.68 亿，互联网普及率达到 48.8%。手机网民保持增长态势，已达 5.94 亿[2]。互联网普及率的快速提升，使得在线社会网络[3]发展极快，以微博为代表的各种社交平台已然成为信息传播的中坚力量，它在带给人们信息传播方便性的同时，也为我国的舆情工作的开展带来了挑战。据不完全统计，腾讯微博与新浪微博目前注册用户总数已达十亿数量级别，日增数据量达到 TB 级别，海量数据的出现以及如何从如此庞大的数据量中进行挖掘、分析，获取重要的信息，实现对敏感信息、热点话题的检测跟踪等舆情监控分析成为一个重要研究方向以及我国舆情工作者面临的巨大挑战。

大数据时代，数据在爆发式增长，然而传统的舆情监控系统大部分是基于工作站或者服务器，使得运营成本很高，传统的数据库方案在海量数据处理方面往往表现为成本高昂、可扩展性差、单点通信故障等。利用 Hadoop 大数据技术处理海量数据成为当下热门的解决方案，因此，本文构建基于 Hadoop 的微博舆情监控系统模型，可以高效地对海量微博数据进行挖掘分析，达到舆情监控的目的，具有现实意义。

2　相关研究

截至 2015 年 12 月 10 日，通过中国知网中国学术文献网络出版总库页面，选择中国学术期刊网络出版总库、中国博士学位论文全文数据库、中国优秀硕士学位论文全文数据库、中国重要会议论文全文数据库和中国重要报纸全文数据库，以“微博+舆情”为关键词，可以检索出 9 397 篇相关文献。然而当调整检索式为“微博+大数据+舆情”时，仅有 22 篇相关文献。从检索结果来看，大部分学者的研究集中在微博信息的传播方向上，探讨微博信息传播的特点、影响机制等问题。其中，兰月新等[4]通过构建数学模型，研究大数据背景下微博与其他网络媒体的信息交互问题，也有学者根据复杂网络理论对微博信息传播特征进行分析，如田占伟等[5]利用复杂网络理论方法，对构建的微博信息传播网络，进行基于度、路径统计指标的分析，最终表明信息在微博网络中的传播效率比其他在线社会网络更高等特征；同时，也有学者从用户角度出发，

通讯作者：杨爱东，ORCID：0000-0003-0186-6773，E-mail：yangaidongcumt@163.com。

研究微博意见领袖相关问题，如刘志明等[6]从用户影响力和用户活跃度两个角度考虑，构建微博意见领袖指标体系，提出使用层次分析法和粗糙集决策分析理论对意见领袖的特征进行识别及分析的理论框架。在微博舆情监控方面，高承实等[7]提出可利用新浪微博现有的排名功能，对于受众的监测，可以分析受众地区分布，加之受众情绪评估，重点对事件发生地区的稳定度进行实时监测；马彦[8]通过分析大数据环境下微博舆情的发展特点和舆情自动监测的具体需求，设计微博舆情热点挖掘系统结构模型，描述各层的主要功能和实现方法；也有学者根据神经网络进行舆情的研究，如潘芳等[9]构建基于 BP 神经网络的预警监控模型以应对动态多变的微博网络社群突发舆情。目前，鲜有学者根据当前的大数据环境，构建微博舆情监控系统模型，对海量数据进行处理分析，达到预警监控的目的。

在国内，新浪微博是最大的在线社会网络，在微博领域也最具有影响力，是微博领域的代表，因此本文数据源取自新浪微博，在此基础之上，主要介绍微博舆情监控系统模型框架和微博舆情监控系统结构，并进行系统模型的仿真。

3　系统整体框架

Hadoop[10]是 Apache 软件基金会旗下的一个开源项目，由 Apache 软件基金会于 2005 年设计，作为较早出现的云计算平台，其开源特性使得 Hadoop 发展迅速，目前已经拥有成熟的社区，技术上也比较成熟，在数据处理效率、稳定性和容错性方面表现很好，基于以上特点，Hadoop 平台使用者可以自由地开发、运行基于海量数据的应用程序，在性能提高的同时，开发成本也大幅度降低，在当下 Hadoop 被认为是大数据处理的标准。整个平台包括 Hadoop 内核、HDFS[11] (Hadoop 分布式文件系统)、MapReduce[12]并行计算框架以及一些相关开源项目，如 Hive 数据仓库基础架构、HBase[13]非关系型分布式数据库等。

本文微博舆情监控系统模型基于 Hadoop 平台，以 HBase 作为海量数据存储数据库，整个模型包括 Hadoop 基础架构、微博数据采集模块、数据预处理模块、微博舆情监控分析模块以及可视化交互 5 部分，如图 1 所示。

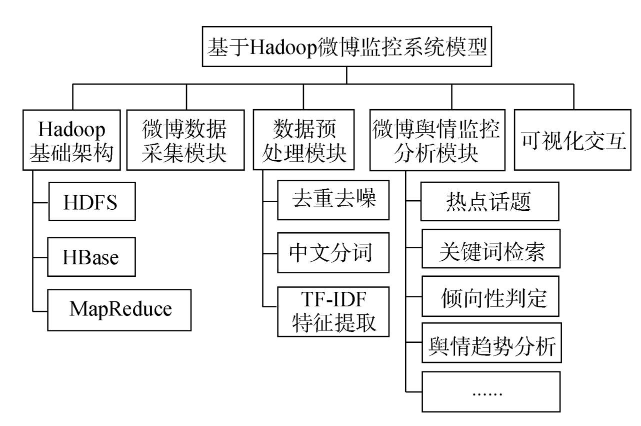


图 1　微博舆情监控系统模型整体框架

4　微博舆情监控系统功能结构

4.1　功能模块分析

微博舆情监控系统由 Hadoop 基础架构、微博数据采集模块、数据预处理模块、微博舆情监控分析模块、可视化交互模块组成。图 2 为微博舆情监控系统各模块交互图。

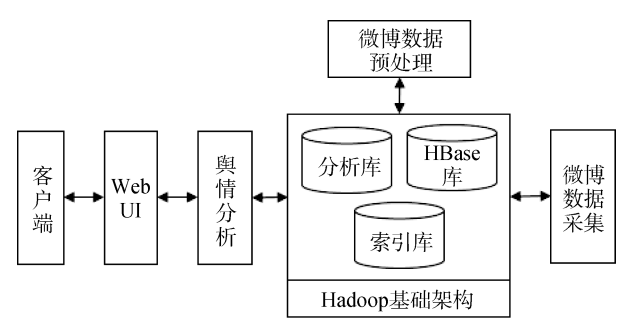


图 2　微博舆情监控系统各模块交互

(1) Hadoop 基础架构：提供 Hadoop 分布式数据(索引库、HBase 库、分析库)的操作接口、MapReduce 并行计算框架；

(2) 微博数据采集模块：采集微博博主相关信息、微博内容、点赞数、转发数、原文链接等信息；

(3) 数据预处理模块：完成数据的去重去噪、中文分词、特征提取等相关工作，为监控分析作数据准备；

(4) 微博舆情监控分析模块：文本的向量化表示、对预处理后的数据进行聚类分析、文本相似度计算等实现舆情监控分析功能；

(5) 可视化交互模块：基于 J2EE 架构的用户交互功能。

4.2　微博数据采集模块关键技术

数据采集模块主要负责新浪微博数据的采集，包

括博主信息、微博内容以及关注信息。获取新浪微博数据主要利用新浪微博服务商提供的 API 接口，使用 API 接口的好处在于方便，并且效率较高，但是在实验过程中发现，新浪微博服务提供商并没有把所有的接口都展现给普通用户，同时对于不同的 API 接口调用的频率与查询范围也进行限制。新浪微博限制了对服务器的一次请求返回的结果数和普通授权用户每小时接口的访问次数，而且拒绝短时间内高频率的 API 接口调用，所以在采集过程中，笔者进行优化，利用队列及轮换使用多个微博账号解决这一问题，得到 JSON 数据后，需要进行解析将数据去噪同时作消重处理，最终存入 Hbase。图 3 为通过新浪微博 API 数据采集流程。

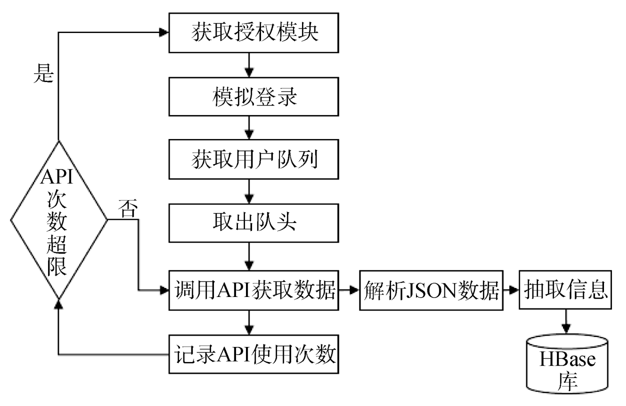


图 3　新浪微博 API 数据抓取程序流程

4.3　数据预处理模块

数据预处理模块主要包括文本的去重去噪、中文分词生成倒排索引文件和文本特征提取三个部分，下面重点介绍通过中文分词得到倒排索引文件和通过特征提取得到文本向量集。整体流程如图 4 所示:

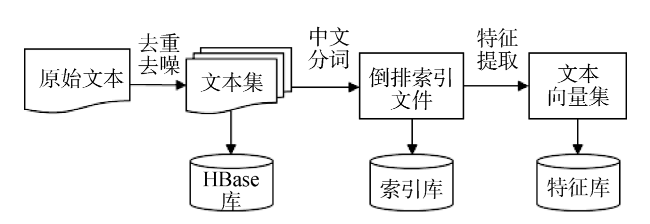


图 4　数据预处理模块流程

(1) 分布式预处理

中文词法分析是中文信息处理的关键和基础。分布式预处理主要使用中国科学院计算技术研究所研发的汉语词法分析系统 ICTCLAS①对文本分词，最终生成倒排索引文件。ICTCLAS 主要功能包括中文分词、词性标注、命名实体识别、新词识别，同时支持用户词典[14]。另外，其对于中文信息的分词性能和分词精度均非常高，分词效果非常好，结合 Hadoop 使用性能比较乐观。

在此阶段，结合 MapReduce 利用 ICTCLAS 分词系统，实现中文分词等功能。Map 阶段主要负责把一行文本 Map 成若干组键值对，在并行的 Reduce 阶段，确保所有经过映射的键值对根据键值的不同共享在同一个组内。图 5 为整体流程图。

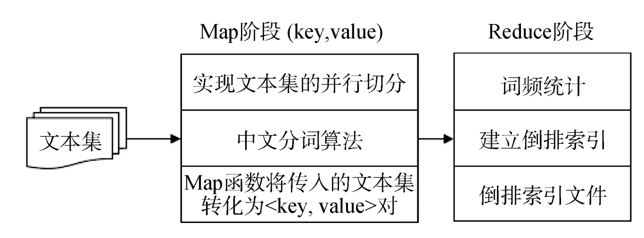


图 5　中文分词系统生成倒排索引文件流程

①将去重去噪的文本，利用 MapReduce 框架，在 Map 阶段实现文本集的并行切分，与传统的 ICTCLAS 分词相比，在这个阶段引入<key, value>对，其中 key 表示经过 Map 函数处理后的‘词’, value 表示经过 Map 函数分词后的‘词频’;

②利用 Reduce 函数进行相同词的汇总，此时的 value 表示进行分词后的词组汇总后的总词频;

③将<key, value>键值对进行对调，即词频在前，词在后，以实现词频按照降序排列。

(2) 特征抽取模块

特征选择的任务是将文本预处理后得到的倒排索引文件进行特征降维处理，计算特征词在各个文本中的权重，最终得到文本向量集合。本文选择 TF-IDF 算法，并在 MapReduce 下进行实现。TF-IDF 是一种基于向量空间模型的分类算法，用以评估一个字或词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降[15]。其优势在于向量模型结构简单、便捷，随着数据规模的增大，分类精度会大幅提高，其性能相当好，并且这种模型易于并行化，非常符合 Hadoop 的核

① http://ictclas.nlpir.org/.

心思想：任务的分割和并行运行[12]。在 Hadoop 分布式平台下可以有效地将文本分类。TF-IDF 算法的处理流程如下：

①在 Map 阶段各个 Mappers 读取索引文件中的文本块；

②统计文档个数和每篇文档中特征词的出现次数，以键值对形式输出；

③将键值对按键的大小进行本地排序后发送给 Reducer，将拥有同一文档 ID 的所有特征词的 TF-IDF 值进行归一化处理；

④将各个特征词的 TF-IDF 值作为文本向量中的项来构建新的文本向量。

特征抽取模块流程如图 6 所示：

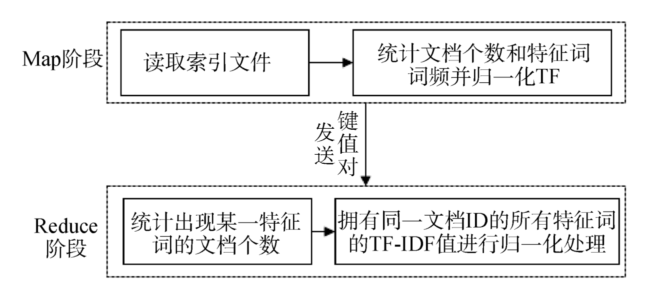


图 6 特征抽取模块流程

4.4 微博舆情监控分析模块

舆情监控分析模块是系统的核心模块，它包含最新消息、热点话题发现、敏感话题检测、话题追踪、情感倾向分析、舆情走势分析、活跃博主追踪等。以下仅对主要功能展开具体的阐述。

(1) 热点话题发现

热点话题发现是网络舆情分析的重点，是在上述构造出的特征矩阵基础上进行文本聚类，利用文本聚类算法计算出相似内容，将聚类后的各个中心点及其子项进行存储，将聚类结果进行可视化输出。在文本聚类过程中，本文结合实际情况针对 K-means 聚类算法进行以下优化：

①由于是处理中文文本，针对汉语一词多义、同义词等情况，在聚类过程中，计算向量乘积时结合 HowNet[16]计算文本相似度，提高聚类的精度；

②由于 K-means 聚类算法对 K 值的变化比较敏感，本文采用 Canopy 算法确定簇数 K 和簇中心；

③运行在 Hadoop 框架上，加快文本处理速度，实现 K-means 聚类算法的并行化。

通过优化后的 K-means 聚类算法并行化方式，达到改善聚类效果，提高聚类精度的目的。工作流程如下：

①读取特征提取模块得到的特征矩阵；

②通过基于 MapReduce 的 Canopy 算法[17]获取簇中心；

③通过优化后的 K-means 算法计算数据对象与簇中心的距离；

④将聚类结果中各中心点以及包含的子项写入分析库，并进行可视化输出。

(2) 情感倾向性分析

微博情感倾向性分析就是对说话人的态度(或称观点、情感)进行分析，也就是对文本中的主观性信息进行分析[18]。情感倾向性分析主要完成的工作简单来说就是利用计算机通过信息发布者的内容自动对文本表达的情感倾向进行判断，将文本感情色彩分为正面褒义类、中立类、负面贬义类三种。

本文采用文献[19]提出的微博情感倾向算法实现文本的倾向性分析，该算法主要是在 Shen[20]提出的 MBEWC 微博情感倾向计算器的基础上，针对微博文本信息的特殊性提出改进算法，加入多种词典并构建全新的情感倾向词典方案，使得情感倾向判别准确率大幅提升。本算法的主要实现流程如图 7 所示：

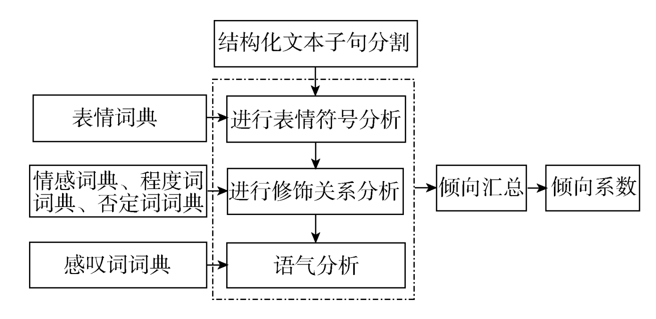


图 7 微博情感倾向算法流程

该过程分别采用 n 个 Map 阶段和一个 Reduce 阶段，将计算结果存储在分析库，通过用户交互模块进行可视化之后展现出来。

5 模型仿真

5.1 实验环境配置

实验在硬件方面由 4 台同构的普通 PC 机通过一台交换机相连构建一个小型的 Hadoop 集群，分别在 4 台 PC 机上安装 14.04 Ubuntu 操作系统，将 Hadoop 2.2.0 部署在 Ubuntu 系统上完成 Hadoop 集群的搭建，将其中一台 PC 机作为主节点，命名为 HostMaster，用来启动 JobTracker 和 NameNode 进程，剩下的三台机器分别命名为 slave1、slave2、slave3 作为从节点，用来启动 TaskTracker 和 DataNode 进程。4 台 PC 机 IP

地址分别为 172.30.78.1- 172.30.78.4。实验的软硬件具体配置及节点拓扑结构如表 1、表 2 和图 8 所示:

表 1　硬件环境配置

硬件	配置
CPU	Intel(R) Core (TM) i3-3240 3.40GHz
硬盘	500GB
内存	4 GB
以太网卡	Realtek PCIe GBE Family Controller
交换机	Mbps

表 2　软件环境配置

软件	软件版本
操作系统	Ubuntu 14.04
JDK	jdk1.7.0_51
Hadoop	Hadoop-2.2.0
HBase	HBase0.96
Eclipse	eclipse-jee-kepler-SR1-linux-gtk-x86_64

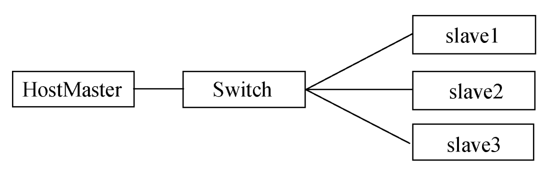


图 8　节点拓扑结构

5.2　数据采集

微博数据采集模块是整个系统数据的来源，它的成功与否关系到后面系统每个环节的实现。采集数据时，不可能采集到新浪微博所有的数据，主要采取广度优先遍历用户列表的策略：从一个受关注度高的种子用户出发，获取其关注用户，形成第一层用户，获取第一层用户的关注列表，形成第二层用户，通过不断向其所关注的微博用户进行扩张的方式，直到用户层数或者本层用户数达到设定的值为止，利用数据采集模块的算法获取到相关的数据。在本次采集中，笔者将"今日头条"作为种子用户，采集的数据时间设定在 2015 年 6 月 1 日–2015 年 11 月 30 日，最终采集接近 15 万微博数据，主要包括微博链接、内容、博主相关个人信息、粉丝相关信息、微博转发数和评论数等。

5.3　文本预处理

在数据采集完成之后，为了方便验证，选取 300MB 数据进行文本预处理实验，主要为了评估在 Hadoop 平台下，系统进行中文分词、特征抽取等文本预处理过程时系统处理效率。一般利用指标加速比衡量不同节点下的系统的性能。加速比(Speedup)是同一个任务在单处理器系统和并行处理器系统中运行消耗的时间的比率，衡量并行系统或程序并行化的性能和效果，加速比计算公式如下:

$$\text{Speedup} = \frac{T1}{Tp} \qquad (1)$$

其中，T1 是单处理器(单节点)下的运行时间，Tp 是在有 P 个处理器(多节点)并行系统中的运行时间。

(1) 中文分词实现倒排索引文件

该实验主要是为了评估在分布式环境下进行分词、倒排索引的构建等文本预处理过程时系统处理效率，分别在节点数目为 1、2、3 时，得到单机系统和集群规模下采用不同节点个数进行实验所耗费的时间，实验结果如表 3 所示:

表 3　中文分词处理实现倒排索引文件时间及加速比

节点个数	时间(s)	加速比
单机	1 268.9	–
1	1 369.4	0.93
2	685.6	1.85
3	513.7	2.47

(2) 文本向量化

使用特征选择后得到的特征词对文本进行向量化处理，计算节点数目不同时的加速比，实验结果如表 4 所示:

表 4　文本向量化处理时间及加速比

节点个数	时间(s)	加速比
单机	448.5	–
1	492.8	0.91
2	285.7	1.57
3	242.4	1.85

对中文分词实现倒排索引文件、文本向量化两个阶段绘制加速比曲线，如图 9 所示:

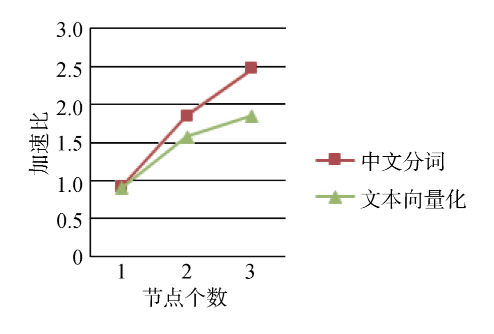


图 9　文本预处理实验加速比对比

(3) 实验结果分析

①在单节点处理时，其加速比略小于 1，这是由于节点自身的 TaskTracker 和 DataNode 上的进程会有通信开销，处理速度比单机系统要差，但是影响不是很大，这一特点在实验的各个阶段都有反映。

②随着节点数目的增加，Hadoop 的性能优势也显现出来：加速比增加，并且加速比越来越大，各个阶段系统开始并行运行。中文分词、文本向量化的处理速度明显提升，说明节点越多，数据块的分割粒度越细，任务运行的并发程度就会越高。

③加速比并不与节点个数成正比增长，与正比增长相比，会稍有所降低，原因在于节点之间相互通信所耗费的时间增加，从而影响了并行效率。由于节点数目较少，如果继续增加节点数目，可以更清晰看到这个特点。

5.4　热点话题发现及可视化

在热点话题发现这一阶段，利用采集到的数据，从中提取出博主 ID、发布时间、微博内容、采集时间等相关字段，通过中文分词、特征抽取、向量化文本，以余弦相似性度量对微博数据进行聚类，将 2015 年 8 月 15 日当天的数据聚类后的结果进行可视化，结果如图 10 所示，中心的“+”表示当日的话题，周围的符号表示参与此话题的博主。

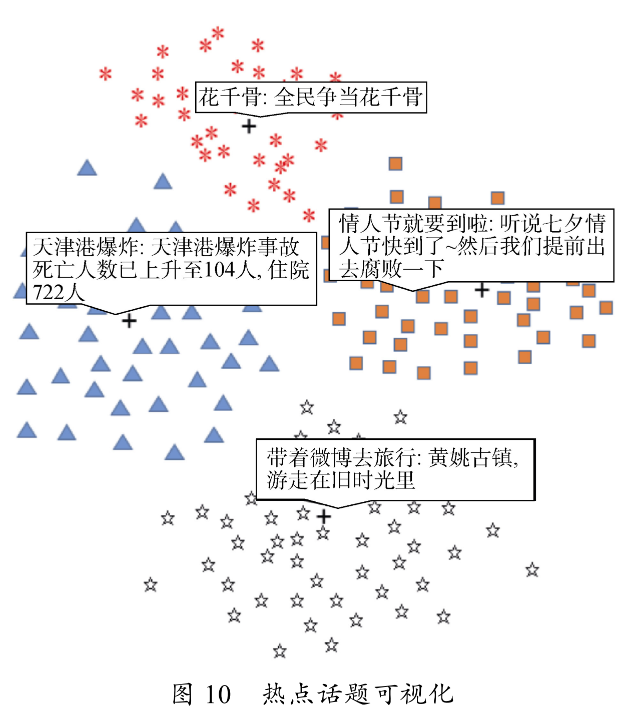


图 10　热点话题可视化

5.5　情感倾向性分析

在情感倾向性分析阶段，为了便于统计与计算，随机抽取了一万条的微博数据进行情感倾向性的判定，并对其进行人工的标注判定。在标注过程中，主要采用以下流程对获取到的数据进行判定：由 5 人完成标定，每人判定 2 000 条微博数据的倾向性，以加快数据的判定；每人标注完成后，将微博内容相同但是标定为不同情感倾向的微博进行讨论，作出统一判定标准；对不同意见的微博进行讨论，以少数服从多数的原则进行标定，直到将所有数据完成标注；最终统计三类微博倾向数量如表 5 所示：

表 5　人工标注判定微博倾向统计

微博数据总量	积极倾向	中立倾向	消极倾向
10 000	3 625	4 617	1 758

利用文献[19]提出的算法计算准确率(Precision)与召回率(Recall)。准确率用以评估算法的准确度，召回率用以评估该算法识别出原来具有某种倾向的微博文本被成功识别的概率。准确率与召回率计算公式如下所示：

$$\text{Precision}=\frac{\text{Correct}}{\text{Propose}}\times 100\% \tag{2}$$

$$\text{Recall}=\frac{\text{Correct}}{\text{Gold}}\times 100\% \tag{3}$$

其中，Correct 指分类正确的数量，Propose 指所提交结果中认为是该分类的数量，Gold 为样本中人工标记的该分类的数量。表 6 和表 7 分别为准确率和召回率计算后的统计结果，从整体表现来看，该算法的准确率较高，成功完成了微博倾向性的自动判定，对实际舆情工作有一定的指导意义。

表 6　准确率统计结果

微博数据	积极倾向	中立倾向	消极倾向
Propose	3 665	4 379	1 956
Correct	2 750	3 829	1 379
Precision	75.03%	87.44%	70.50%

表 7　召回率统计结果

微博数据	积极倾向	中立倾向	消极倾向
Gold	3 625	4 617	1 758
Correct	2 750	3 829	1 379
Recall	75.86%	82.93%	78.44%

6　结　语

本文针对微博这一社交网络的快速发展，提出基

于 Hadoop 的微博舆情监控系统模型，研究大数据环境下，将 Hadoop 分布式存储和 MapReduce 并行计算模型运用于海量微博舆情监控分析，并对模型组成模块的工作流程和实现方式做了详细设计。本文主要完成以下工作:

(1) 研究网络舆情分析的关键技术，深入分析信息采集、信息预处理、文本聚类等各个模块，完成整个模型框架的构建;

(2) 利用普通PC机构造 Hadoop 集群，对提出的模型在不同的节点下的系统性能进行对比分析;

(3) 完成数据的抓取工作，并利用文献[19]提出的算法，成功将抓取到微博进行情感倾向性的判定;

(4) 对提出的基于Hadoop的微博舆情监控系统模型进行验证。

通过实验仿真，基于 Hadoop 的微博舆情监控系统可以有效地对大规模微博数据进行舆情监控分析，然而仍存在以下问题需要进行后续研究:

(1) 对实验条件进行改进，扩大 Hadoop 集群，尝试更多节点下模型的效率;

(2) 尝试其他聚类算法，进行对比分析，完成基于 Hadoop 平台下微博舆情监控系统获取热点话题的准确度;

(3) 本文微博舆情监控系统主要研究工作集中处理微博文本，在后续工作中要多注重多媒体数据的处理，以获取更大的实用价值。

参考文献:

[1] 张克生. 国家决策: 机制与舆情[M]. 天津: 天津社会科学院出版社, 2004: 17. (Zhang Kesheng. National Decision-making: Mechanism and Public Opinion [M]. Tianjin: Tianjin Academy of Social Sciences Press, 2014: 17.)

[2] 中国互联网络信息中心. 第36次中国互联网络发展状况统计报告[R/OL]. [2015-07-23]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201507/P020150723549500667087.pdf. (CNNIC.The 36th China Internet Development Statistics Report [R/OL]. [2015-07-23]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201507/P020150723549500667087.pdf.)

[3] Wasserman S, Faust K. Social Network Analysis: Methods and Applications [M]. Cambridge, NY: Cambridge University Press, 1994.

[4] 兰月新, 董希琳, 苏国强, 等. 大数据背景下微博舆情信息交互模型研究[J]. 现代图书情报技术, 2015(5): 24-33. (Lan Yuexin, Dong Xilin, Su Guoqiang, et al. Research on Micro-blog Public Opinion Information Interaction Model Under the Background of Big Data [J]. New Technology of Library and Information Service, 2015(5): 24-33.)

[5] 田占伟, 隋玚. 基于复杂网络理论的微博信息传播实证分析[J]. 图书情报工作, 2012, 56(8): 42-46. (Tian Zhanwei, Sui Yang. The Empirical Analysis of Micro-blog Information Flow Based on Complex Network Theory [J]. Library and Information Service, 2012, 56(8): 42-46.)

[6] 刘志明, 刘鲁. 微博网络舆情中的意见领袖识别及分析[J]. 系统工程, 2011, 29(6): 8-16. (Liu Zhiming, Liu Lu. Recognition and Analysis of Opinion Leaders in Microblog Public Opinions [J]. Systems Engineering, 2011, 29(6): 8-16.)

[7] 高承实, 荣星, 陈越. 微博舆情监测指标体系研究[J]. 情报杂志, 2011, 30(9): 66-70. (Gao Chengshi, Rong Xing, Chen Yue. Research on Public Opinion Monitoring Index-system in Micro-blogging [J]. Journal of Information, 2011, 30(9): 66-70.)

[8] 马彦. 大数据环境下微博舆情热点话题挖掘方法研究[J]. 现代情报, 2014, 34(11): 29-33. (Ma Yan. Study on the Method of Micro-blogging Public Opinion Hotspots Mining in Big Data [J]. Modern Information, 2014, 34(11): 29-33.)

[9] 潘芳, 张霞, 仲伟俊. 基于BP神经网络的微博网络社群突发舆情的预警监控[J]. 情报杂志, 2014, 33(5): 125-128. (Pan Fang, Zhang Xia, Zhong Weijun. Precautionary Monitoring of the Sudden Burst of Public Opinion in Weibo Community on Internet Based on BP Neural Network [J]. Journal of Information, 2014, 33(5): 125-128.)

[10] Hadoop [EB/OL]. [2016-01-12]. http://hadoop.apache.org/.

[11] HDFS User Guide [EB/OL]. [2016-01-12]. http://hadoop.apache.org/docs/r1.2.1/hdfs_user_guide.html.

[12] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters [J]. Communications of the ACM, 2004, 51(1): 107-113.

[13] George L. HBase: The Definitive Guide [M]. O'Reilly Media, 2011.

[14] Song Y, Cai D F, Zhang G P, et al. Approach to Chinese Word Segmentation Based on Character-Word Joint Decoding [J]. Journal of Software, 2009, 20(9): 2366-2375.

[15] TF-IDF [EB/OL]. [2016-01-12]. http://baike.baidu.com/view/1228847.htm.

[16] 知网 [EB/OL]. [2016-01-12]. http://www.keenage.com/. (HowNet Knowledge Database [EB/OL]. [2016-01-12]. http://www.keenage.com/.)

[17] 李应安. 基于MapReduce的聚类算法的并行化研究[D]. 广州:中山大学, 2010. (Li Ying'an. Research on Parallelization of Clustering Algorithm Based on MapReduce [D]. Guangzhou: Sun Yat-Sen University, 2010.)

[18] 冯希莹, 王来华. 舆情概念辨析[J]. 社会工作, 2011(10): 83-87. (Feng Xiying, Wang Laihua. Discussion of the Concept of Public Opinion and Sentiments [J]. Journal of Social Work, 2011(10): 83-87.)

[19] 张伟舒, 吕云翔. 微博情感倾向算法的改进与实现[J]. 知识管理论坛, 2013(9): 21-27. (Zhang Weishu, Lv Yunxiang. The Improvement and Implementation of the Micro-blog Sentiment Orientation Algorithm [J]. Knowledge Management Forum, 2013(9): 21-27.)

[20] Shen Y. Emotion Mining Research on Micro-blog [C]. In: Proceedings of the 1st IEEE Symposium on Web Society. Lanzhou: Lanzhou University, 2009.

作者贡献声明:

刘东苏: 提出研究思路, 设计研究方案, 论文最终版本修订;
杨爱东: 设计实验, 实验数据采集、预处理和分析, 论文撰写。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: yangaidongcumt@163.com。

[1] 杨爱东, 刘东苏. blog_data.rar. 抓取的自 2015 年 6 月 1 日–2015 年 11 月 30 日的新浪微博数据.

[2] 杨爱东, 刘东苏.order_index.rar. 由中文分词数据预处理后得到的倒排索引文件.

[3] 杨爱东, 刘东苏. attr_reduce.rar. 由特征选择后得到的文本向量集合.

[4] 杨爱东, 刘东苏. topic.rar. 热点话题发现阶段使用的预处理后的微博数据.

[5] 杨爱东, 刘东苏. data_label.rar. 对一万条微博数据经过人工标注后的数据.

收稿日期: 2015-12-11
收修改稿日期: 2016-01-29

Hadoop Based Public Opinion Monitoring System for Micro-blogs

Yang Aidong　Liu Dongsu
(School of Economics and Management, Xidian University, Xi'an 710126, China)

Abstract: [**Objective**] This paper presents a new model for public opinion monitoring system based on Hadoop to retrieve and analyze information from the micro-blog platforms. [**Methods**] We first surveyed the existing technology of the public opinion monitoring systems and proposed a new model with modified algorithm. Then, we built a big data analysis platform with Hadoop to examine the model's feasibility through experimental simulations. [**Results**] The proposed model can detect and retrieve public opinion data effectively. [**Limitations**] The Hadoop cluster was relatively small. We did not compare our model with other clustering algorithms to discuss their advantages and disadvantages. [**Conclusions**] The proposed model can conduct public opinion analysis with micro-blog data and provide scientific information for the policy makers to improve crisis management.

Keywords: Monitoring public opinion　Hadoop　Micro-blog　Big data